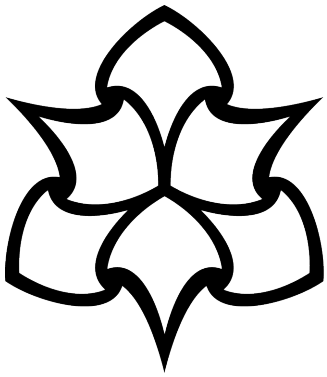


Reviews in Fisheries Science & Aquaculture

ISSN: 2330-8249 (Print) 2330-8257 (Online) Journal homepage: http://www.tandfonline.com/loi/brfs21

Using Machine Learning to Uncover Latent Research Topics in Fishery Models

Shaheen Syed & Charlotte Teresa Weber

To cite this article: Shaheen Syed & Charlotte Teresa Weber (2018) Using Machine Learning to Uncover Latent Research Topics in Fishery Models, Reviews in Fisheries Science & Aquaculture, 26:3, 319-336, DOI: 10.1080/23308249.2017.1416331

To link to this article: https://doi.org/10.1080/23308249.2017.1416331

REVIEWS IN FISHERIES SCIENCE & AQUACULTURE
2018, VOL. 26, NO. 3, 319–336
https://doi.org/10.1080/23308249.2017.1416331

Taylor & Francis
Taylor & Francis Group

Using Machine Learning to Uncover Latent Research Topics in Fishery Models

Shaheen Syed [a,b] and Charlotte Teresa Weber [c]

[a]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands; [b]Centre for Policy Modelling, Manchester Metropolitan University, All Saints Campus, Manchester, United Kingdom; [c]Norwegian College of Fishery Science, UiT – The Arctic University of Norway, Tromsø, Norway

ABSTRACT

Modeling has become the most commonly used method in fisheries science, with numerous types of models and approaches available today. The large variety of models and the overwhelming amount of scientific literature published yearly can make it difficult to effectively access and use the output of fisheries modeling publications. In particular, the underlying topic of an article cannot always be detected using keyword searches. As a consequence, identifying the developments and trends within fisheries modeling research can be challenging and time-consuming. This paper utilizes a machine learning algorithm to uncover hidden topics and subtopics from peer-reviewed fisheries modeling publications and identifies temporal trends using 22,236 full-text articles extracted from 13 top-tier fisheries journals from 1990 to 2016. Two modeling topics were discovered: estimation models (a topic that contains the idea of catch, effort, and abundance estimation) and stock assessment models (a topic on the assessment of the current state of a fishery and future projections of fish stock responses and management effects). The underlying modeling subtopics show a change in the research focus of modeling publications over the last 26 years.

KEYWORDS

Topic models; latent Dirichlet allocation; fisheries science; fisheries models; research trends

1. Introduction

Global research efforts have increased significantly in recent years (Oecd, 2008), as has publication output within fisheries science (Aksnes and Browman, 2016). This growth has been partly driven by growing concerns about the state of fish stocks and the need to provide information for policy and decision makers globally. Since each fish stock is typically unique, and experimental approaches cannot be used to predict their response to fishing, it follows that the modeling and simulation of fisheries play a major role in providing management advice; these are among the most frequently used methods in fisheries science (Jarić et al., 2012). Models offer a feasible approach to the approximation of trends and processes, and they advance the understanding of fisheries and ecosystem dynamics (Angelini and Moloney, 2007) while guiding data collection and illuminating core uncertainties (Epstein, 2008). For this reason, and in contrast to common perceptions, a multitude of fisheries models is available besides standard stock assessment models, and these models take on many different shapes and forms depending on their method and purpose. Such models may include individual-based models to investigate fleet behavior (Bastardie et al., 2014); Bayesian belief networks to better understand stakeholder viewpoints and perceptions (Haapasaari et al., 2012); or conceptual models to analyze fisheries from a socio-ecological complex adaptive system perspective (Ostrom, 2009; Partelow, 2015).

The frequent use of models and their wide range of applications, in combination with the growing global collections of scholarly literature, have led to an ever-increasing number of publications on the various types of models and approaches. As a result, scientists are suddenly faced with millions of publications, overwhelming their capacity to effectively use these collections and to keep track of new research (Larsen and von Ins, 2010). Online collections can be browsed and explored using keyword searches, through which publications can be collected manually; however, in addition to being time-consuming, the size and growth of the body of research often has the effect of limiting the possibility of identifying all the relevant literature. Another problem is that the underlying topic of an article is not readily available in most collections. Thus, the topic of an article – that is, the idea underlying the article, which may be shared with similar articles – cannot always be detected using

CONTACT Shaheen Syed s.a.s.syed@uu.nl Centre for Policy Modelling, Manchester Metropolitan University, All Saints Campus, Oxford Road - Manchester M15 6BH, UK.

keyword searches (Srivastava and Sahami, 2009). Given such challenges, an assessment of the field of fisheries models could reveal overlooked research topics, identify important changes in research directions (i.e., trends), assess the diversity of topics in publication outlets, and ultimately help in identifying new and emerging modeling topics. Furthermore, an improved understanding of fisheries modeling approaches could help researchers to more easily synthesize historical and current research developments.

The developments and trends in fisheries science and fishery models are usually assessed through reviews (e.g., Bjørndal et al., 2004; Prellezo et al., 2012) and bibliometric studies (Jarić et al., 2012; Aksnes and Browman, 2016). These types of studies have several limitations, such as taking into account only a limited number of publications (e.g., only 61 publications, Gerl et al., 2016); a limited time period (e.g., from 2000 to 2009, Jarić et al., 2012); a limited scope or very specialized focus (e.g., stock assessment methods, Cadrin and Dickey-Collas, 2015; bio-economic models, Prellezo et al., 2012; models of an ecosystem approach to fisheries, Plagányi, 2007; and models of the Celtic Sea, Minto and Lordan, 2014). Other limitations include proxies for full text such as titles (Jarić et al., 2012) and abstracts (Aksnes and Browman, 2016), and proxies for research topics such as one word per topic (Jarić et al., 2012; Aksnes and Browman, 2016). Most importantly, previous attempts to identify trends in fisheries and fisheries modeling are based on top-down approaches, in which research topics are predefined by the researcher (Debortoli et al., 2016), such as region, species, habitat, or study area. Such approaches are prone to human subjectivity; researchers may end up with different results (Urquhart, 2001), or the mapping of text features to categories may not be explicitly known (Quinn et al., 2010).

This study aims to overcome the limitations of previous approaches by applying a bottom-up approach in which research topics automatically emerge from the statistical properties of the documents. In doing so, the topics are automatically uncovered without prior human labeling, categorization, or predefined classification of publications, and they are thus not biased by researchers' top-down subjective choices. For this purpose, a probabilistic topic model algorithm called latent Dirichlet allocation (LDA) (Blei et al., 2003), which belongs to the field of unsupervized machine learning algorithms, was used to reveal research topics within the field of fisheries models that are published in peer-reviewed journals and have a strong focus on fisheries. Topic model algorithms can automatically uncover hidden or latent thematic structures (i.e., topics) from large collections of documents. The unsupervized nature of LDA allows documents to "speak" for themselves, and topics emerge without human intervention. They have proven to be very useful in automatically identifying and interpreting scientific themes in relation to the journal's existing themes or categories (Griffiths and Steyvers, 2004).

By utilizing unsupervized machine learning, this study aims to provide comprehensive information on topical trends within fisheries modeling research for fisheries scientists and stakeholders. In particular, this study analyzes 22,236 full-text scientific publications published within the period from 1990 to 2016 in 13 top-tier fisheries journals. Thus, a unique dataset for the field of fisheries models was created, and topics in fisheries modeling and their underlying subtopics were identified to determine historical and current research interests. In addition, the species, areas, and methods occurring within the identified topics were assessed.

2. Methods

2.1. Latent Dirichlet allocation

The LDA model is a generative probabilistic topic model that represents documents (i.e., fisheries publications) as discrete distributions over K latent topics; each topic is subsequently represented as a discrete distribution over all the words (i.e., vocabulary) used. The words with high probability within the same topic are frequently co-occurring words, which can be seen as clusters or constellations of words that are often used to describe an underlying topic or theme (DiMaggio et al., 2013). In this way, LDA captures the heterogeneity of research ideas or topics within publications. The topics and their relative proportions within documents are hidden (i.e., latent) variables that LDA infers from the observable variables – that is, the words within the documents. The generative process behind LDA involves an imaginary random process, through which documents are created based on probabilistic sampling rules. The topics and their proportions are subsequently inferred from these generated documents by applying statistical inference techniques, such as variational and sampling-based algorithms (Blei and Jordan, 2006; Teh et al., 2006; Hoffman et al., 2010; Wang et al., 2011). LDA extends other popular topic model algorithms such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) while also overcoming their limitations. An explanation of LDA's generative process can be found in Appendix 1.

The LDA model makes two assumptions when analyzing and uncovering latent topics from documents. First, documents are represented as "bags of words" (i.e., unordered lists of words) in which the

word order is neglected. Although this is an unrealistic assumption, it is reasonable if the aim is to uncover semantic structures from text (Blei and Lafferty, 2006; Blei, 2012). Consider a thought experiment where one imagines shuffling all the words in a document. Even when shuffled, one might find words such as "population," "size," "virtual," "minimum," and "recruitment" and expect that the document deals with aspects of population dynamics. One of the core underlying principles of LDA is based on word co-occurrences, and a small number of co-occurring words is sufficient to resolve problems of ambiguity. Second, LDA assumes that the order in which documents are analyzed is unimportant (i.e., document exchangeability is assumed); however, at the end of the analysis, all documents are analyzed. As a result, LDA is unable to explicitly capture the evolution of topics over decades or centuries of work. This would require a more complicated and computationally expensive dynamic topic model (Blei and Lafferty, 2006), which is currently not feasible given the large dataset; however, this is a potential approach for future work. Document exchangeability is a limitation in the case of topics whose presentation in the literature has dramatically changed (e.g., in terms of the terminology used to describe the topic), but it still captures the phenomenon by which current literature builds upon previous literature. Nonetheless, the assumption of document exchangeability is especially problematic when analysing topics that span 50–100 years of research.

2.2. Topic interpretation

The topics emerge from the statistical properties of the documents and the statistical assumptions behind LDA. The topics are represented as discrete distributions over all the words, in which the top words (e.g., top 15) for each topic – that is, the words with the highest probability and those that more frequently co-occur together – provide insights into the semantic meaning of the topic. Topics are thus a reference to these probability distributions over words to exploit text-oriented intuitions. No epistemological claims are made beyond this representation. Furthermore, by no means is the topic distribution over words limited to these top 15 words; in fact, every word occurs in every topic, but with different probabilities. The topics are used to uncover the themes prevailing the documents, as well as the extent to which such themes are present in each document. In doing so, the main ideas of a publication can be extracted and used to track how they have developed over time. Note that the underlying topics and to what extent the document exhibits these topics are not known in advance. These details are the output of the LDA analysis and emerge automatically from the statistical properties of the documents and the assumptions behind LDA.

2.3. Creating the dataset

This paper aims to identify latent fisheries modeling topics from scientific research articles published in peer-reviewed journals specializing in fisheries. In this manner, the selection of publications was restricted exclusively to fisheries journals; therefore, it follows that some subjective choices were made to achieve this. All journals included in this analysis contain the term "fishery" or "fisheries" in their title and have an impact factor of 1.0 or higher. Additionally, the journal *The ICES Journal of Marine Science* was included, because it is part of the International Council for the Exploration of the Seas (ICES), which channels science-based advice to decision makers for sustainable fisheries, and fisheries models are an important focus of this journal. A total of 13 fisheries journals were included in the study (see Table 1). A time frame of 26 years, from 1990 to 2016, was chosen to allow for enough variation within publication trends. Due to difficulties with journal subscription rights and the fact that some journals started after 1990 (e.g., *Fish and Fisheries* was first published in 2000), coverage was incomplete for the complete time range of 26 years for a few journals. Documents that did not constitute a type of research article (e.g., book reviews, forewords, errata, conference reports, comments, policy notes, corrigenda, and letters) were discarded. In total, 22,236 full-text research articles from 13 top-tier fisheries journals were downloaded using automated download scripts, as well as by utilizing the available application programming interfaces (APIs) offered by the publishers. The use of full-text articles, in contrast to only using abstracts, has shown to increase topic quality and provide a more detailed overview of the latent topics permeating a document collection (Syed and Spruit, 2017). Table 1 provides an overview of the complete dataset utilized in this study.

The selection of fisheries journals and underlying fisheries publications comes with some limitations. First, some of the highly influential and most cited papers on fisheries models are published in high-impact journals such as *Nature, Science*, and *PNAS*. Although highly influential, such publications would constitute only a small number of our sample and would only marginally or even negligibly contribute to the overall number of 22,236 publications downloaded from fisheries journals for this study. Two other reasons exist to exclude such generic journals. The first reason is that including all publications published in such outlets would drastically

Table 1. Overview of the dataset (i.e., corpus): years represent the years for which documents (i.e., articles) are downloaded; IF, the journal's impact factor according to ISI Journal Citation Reports 2016; *N*, the number of documents; *N/T*, the percentage of journal articles in relation to the total number of articles; $\overline{W}$, the mean number of words within each document; Std. W, the estimated standard deviation of words within each document; and $\overline{V}$, the mean vocabulary size (number of unique words) within each document. The total number of documents is 22,236.

Journal	Years	IF	*N*	*N/T*	$\overline{W}$	Std.W	$\overline{V}$
Canadian Journal of Fisheries and Aquatic Sciences	1996–2016	2.44	4427	19.9%	4075.5	1305.5	1266.7
Fish and Fisheries	2000–2016	8.26	419	1.9%	5892.9	2801.4	1757.4
Fisheries	1997–2016	2.43	477	2.1%	3409.9	1633.2	1312.3
Fisheries Management and Ecology	1994–2016	1.51	1001	4.5%	2692.2	1135.7	955.5
Fisheries Oceanography	1997–2016	2.73	752	3.4%	3866.7	1353.8	1187.8
Fisheries Research	1995–2016	2.23	3610	16.2%	3204.4	1326.3	1064.4
Fishery Bulletin	1990–2016	1.51	1441	6.5%	3356.3	2037.0	1074.4
ICES Journal of Marine Science	1990–2016	2.63	3903	17.6%	3379.8	1378.7	1118.9
Marine and Coastal Fisheries	2009–2016	1.44	274	1.2%	4473.7	1363.8	1368.0
North American Journal of Fisheries Management	1997–2016	1.01	2517	11.3%	3288.9	1420.9	1036.6
Reviews in Fish Biology and Fisheries	1991–2016	3.22	659	3.0%	5799.8	3994.4	1750.1
Reviews in Fisheries Science & Aquaculture	1997–2016	2.03	375	1.7%	6185.6	6020.2	1737.3
Transactions of the American Fisheries Society	1997–2016	1.47	2381	10.7%	3887.8	1382.4	1202.7
		Total	22,236				

increase the number of uncovered topics, as fisheries make up a small portion of the publications in *Nature, Science* and *PNAS*. While one might be able to use keyword searches and include only those publications that match fisheries-related terms, this brings up the second reason to exclude such journals: publication filtering is based on the subjective choice of relevant keywords and is limited in terms of how publications are indexed and subsequently can be retrieved (e.g., title, abstract, or full text) from these journals. Through the inclusion of publications from only fisheries journals, such subjective choices and associated limitations are avoided.

The second limitation concerns the exclusion of non-fisheries-specialized journals in which fisheries-modeling-related publication might appear. Such journals focus on, but not limited to, the field of marine science (e.g., *Marine Policy* and *Advances in Marine Biology*), the field of coastal areas or zones (e.g., *Coastal Management* and *Ocean and Coastal Management*), the field of toxicology (e.g., *Environmental Toxicology and Pharmacology* and *Aquatic Toxicology*), and the field of modeling (e.g. *Environmental Modelling & Software* and *Ecological Modelling*), in addition to a number of other journals, such as *Developmental Dynamics, Bulletin of the American Meteorological Society, Environmental Science and Technology, Philosophical Transactions of the Royal Society, Environmental Health Perspectives, BioScience, Journal of Fish Biology*, and *Progress in Oceanography*. Some publications related to fisheries modeling approaches are published in these outlets, which is a potential limitation of this study. Again, filtering for fisheries modeling publications in these journals would be biased by the subjective choice of keywords and limitations due to indexing and retrieval functionalities. Consequently, publications with a focus on the novelty in modeling approaches, which are commonly published in specialized modeling journals such as *Ecological Modeling*, were not assessed in this study. On the other hand, the modeling publications captured within the fisheries journals included in this study can potentially address other topics besides fisheries, such as climate change or habitat loss, which are likely to be included in the analysis of modeling publications.

The third limitation relates to the focus on peer-reviewed journals only. As a result, fisheries modeling research that appears in grey literature was excluded. As grey literature is not indexed in the same way as peer-reviewed studies, selecting only relevant grey literature would, again, introduce bias due to human subjectivity in the search and retrieval.

2.4. Preprocessing the dataset

Several important preprocessing steps were required to transform the documents into appropriate bag-of-word representations. First, each document was converted from PDF format into a plain-text representation. Image-based PDFs, mainly old documents from the 1990s, were converted using the Tesseract optical character recognition (OCR) library. Second, documents were tokenized, which involved creating individual words (e.g., from paragraphs and sentences); meanwhile, numbers, single characters, punctuation marks, and words with only a single occurrence were removed, since they bear no topical meaning. Additionally, words that occurred in $\geq$90% of the documents were discarded due to their lack of distinctive topical significance (see Appendix 2). Boilerplate content, such as title pages, article metadata, footnotes, margin notes and so on, was also removed. The reference list of each article was maintained so as to allow for referenced

titles and names of authors to be part of the word distributions of topics. An advantage of this approach is that author names can be part of specific topics, but they can simultaneously introduce bias when the referenced articles have no direct link to the underlying topics. A standard English stop word list ($n = 153$) was used to remove words that serve only syntactical and grammatical purposes, such as *the, and, were*, and *is*. Finally, other than grouping lowercase and uppercase words, no normalization method was applied, such as stemming or lemmatization, to reduce the inflectional and derivational forms of words to a common base form (e.g., *fishing* and *fishery* to *fish*). Normalization reduces the interpretability of topics at later stages, as stemming algorithms can be overly aggressive and may result in unrecognizable words when interpreting topics. Stemming might also lead to another problem, as it cannot be deduced whether a stemmed word comes from a verb or a noun (Evangelopoulos et al., 2012). For these reasons, and considering that the interpretability of the topics at a later stage was considered to be highly significant, an extensive normalization phase was omitted.

2.5. Creating LDA models

The LDA models were created with the Python library Gensim (Rehurek and Sojka, 2010). The number of topics to be uncovered (i.e., K parameter) varied from 1 to 50, thus creating 50 different LDA models. The hyperparameters for the LDA models, which affect the sparsity of the topics created and their relative proportions, were set to be symmetrical. Technically, since LDA is a Bayesian probabilistic model, the symmetrical hyper-parameters encode prior knowledge that a priori assign equal probabilities to topics within documents, and words within topics. The quality of each topic was calculated using a topic coherence measure to find the optimal value for K (analogous to finding the right number of clusters, e.g., K-nearest neighbors). A coherence measure calculates the degree of similarity between a topic's top N words. This provides a quantitative approach for assessing the interpretability of topics from a human perspective. As such, coherence measures aim to find coherent topics – a topic with top words *apple, pear*, and *banana* is more coherent than *apple, pear*, and *car* – rather than topics that are merely artefacts of the statistical assumptions behind LDA. The C_V coherence measure was adopted, since it has shown the highest accuracy of all available coherence measures (Röder et al., 2015). An elbow method was employed to find the K value with the best performing topic coherence score. A detailed description of the C_V coherence measure can be found in Appendix 3.

2.6. Identifying subtopics

For each modeling topic identified, a zoom-in was employed with the aim of uncovering underlying subtopics within each of the general modeling topics by applying an approach similar to that described above. These subtopics provide a more detailed deconstruction of the respective general modeling topics. A zoom-in is performed on a subset of the data consisting of documents that have the general modeling topic as the dominant topic. The dominant topic is defined as the topic with the highest relative proportion – that is, the topic that exceeds all other topic proportions within a document. Since documents are modeled as mixtures of topics, the dominant topic represents the primary topic of a document.

2.7. Labeling the topics

The LDA model outputs the uncovered topics as probability distributions over all the words used; when sorted, the top 15 words are used to label the topic semantically. Representing the words as probabilistic topics has the distinct advantage that each topic is now individually interpretable (Griffiths et al., 2007), compared to a purely spatial representation like the topic model of latent semantic analysis (Deerwester et al., 1990). As stated before, the distributions of words, and specifically the words with the highest probability within each topic, are used to describe an underlying theme; however, such themes are latent, and a semantic label that best captures those words needs to be attached. For example, a topic with the top 5 words *apple, banana, cherry, pear*, and *mango* describes the underlying theme of fruits and can be labeled as such.

To provide a semantically meaningful and logical interpretation of these probability distributions, a fisheries domain expert manually labeled the topics by close inspection of the top 15 high-probability words, together with an inspection of the document titles and content. Furthermore, to improve the labeling of the topics, the topics were visualized in a two-dimensional area by computing the distance between topics (Chuang et al., 2005) and applying multi-dimensional scaling (Sievert and Shirley, 2014). This two-dimensional topic representation aided in identifying similarities between topics and thus similarities between topic labels.

2.8 Calculating subtopical modeling trends

To gain insight into the subtopical temporal dynamics of the modeling subtopics, document topic proportions were aggregated into a composite topic-year proportion.

Such composite values provide insights into the prevalence of a modeling subtopic within a certain year, given all the publications within that year. It furthermore enables the analysis of changing topic proportions over the course of 26 years, as proportions increase or decrease for each subtopic and for each year. Additionally, to obtain insight into increasing and decreasing topical trends, a one-dimensional least square polynomial was fitted for different time intervals. The time intervals chosen were 1990–1995, 1995–2000, 2000–2005, 2005–2010, and 2010–2016, so as to allow for historical comparison. The polynomial coefficient is used as a proxy for the trend and defines the slope of the composite topic-year proportions for a range of years. Coefficients are multiplied by the number of years within each time interval to obtain the change measured in percentage points. Positive values indicate increasing or "hot" topics, and negative values indicate decreasing or "cold" topics. Color coding is used to represent the hot (i.e., red) and cold (i.e., blue) topical trends.

(1) ESTIMATION MODELS

word	prob.
MODEL	.015
ESTIMATES	.014
CATCH	.012
SURVEY	.008
SAMPLING	.008
ESTIMATED	.008
MODELS	.007
ESTIMATE	.007
DISTRIBUTION	.007
ABUNDANCE	.006
MEAN	.006
EFFORT	.006
SAMPLE	.005
METHOD	.005
SIZE	.005

(2) STOCK ASSESSMENT MODELS

word	prob.
MODEL	.024
STOCK	.014
MORTALITY	.014
POPULATION	.012
RECRUITMENT	.011
MODELS	.010
BIOMASS	.007
YEAR	.007
RATE	.007
MANAGEMENT	.007
PARAMETERS	.006
ASSESSMENT	.006
FISHERIES	.006
ESTIMATES	.006
FISHING	.005

Figure 1. The two uncovered fisheries modeling topics (i.e., estimation models and stock assessment models) from the dataset containing 22,236 fisheries publications (1990–2016; 13 journals). The figure displays the topic label (top) and the top 15 high-probability words.

3. Results and discussion

3.1. General modeling topics

The optimal LDA model for the complete corpus (N = 22,236 documents) uncovered 31 general fisheries topics. The calculated coherence scores to obtain the optimal number of topics, referred to as the K parameter, can be found in Appendix 3. Among these general fisheries topics, two topics deal with the aspects of fisheries modeling. The publications dealing with these two modeling topics account for 12% (N = 2761 documents) of the total number of publications. The remaining 29 topics, which relate to other aspects of fisheries research, are listed in Appendix 4. A bibliometric analysis of trends in fisheries science found a higher proportion of publications employing models – around 30%, as estimated from publication titles and abstracts from a dataset containing 695 fisheries-related publications (Jarić et al., 2012). Several reasons can be offered to explain why these two percentages differ, such as the used time range and the selected journals; most importantly, the present paper identifies publications which predominantly deal with fisheries modeling aspects, in contrast to publications in which a modeling method is employed.

Figure 1 shows the top 15 words and their probabilities for the two modeling topics. The first modeling topic concerns catch-effort and abundance estimation methods and is, therefore, given the short name estimation models. It contains the words "catch," "survey," "sampling," "effort," and "sample" among its top 15 words. These words reflect the collection of both fisheries-independent data, which are usually gathered through survey and sampling methods, and fisheries-dependent data (e.g., collected through logbooks), which commonly provide information on catch and effort. These and other obtained data feed into models in order to estimate intermediate parameters such as natural mortality rate or catchability (Hoggarth et al., 2006); this is a phase of research reflected in estimation models through the words "model," "estimates," "estimated," and "estimate." These types of models might also be called retrospective models, since they interpret the past based on collected data.

The second modeling topic concerns modeling approaches for the assessment of the current state of a fishery and future projections and is assigned the short name "stock assessment models." It contains the words "stock," "mortality," "biomass," "rate," and "estimate," which reflect the most commonly used indicators (i.e., fish catch, stock biomass, stock size, and fishing mortality; Hoggarth et al., 2006) to measure the status of the fishery and the state of the stock (Le Gallic, 2002). These indicators link to reference points, which give quantitative meaning to the goals and objectives set for a fishery (Jennings, 2005). Reference points are usually estimated through models that use stock and recruitment data, which is reflected in the words "stock," "population," "recruitment," "management," "parameters," and "estimates" in stock assessment models. Together, indicators and reference points play a crucial role in fisheries management and can be used to give quantitative meanings to the objectives of a fishery (Hoggarth et al., 2006).

The distinction between these two topics shows how they are treated separately in fisheries research

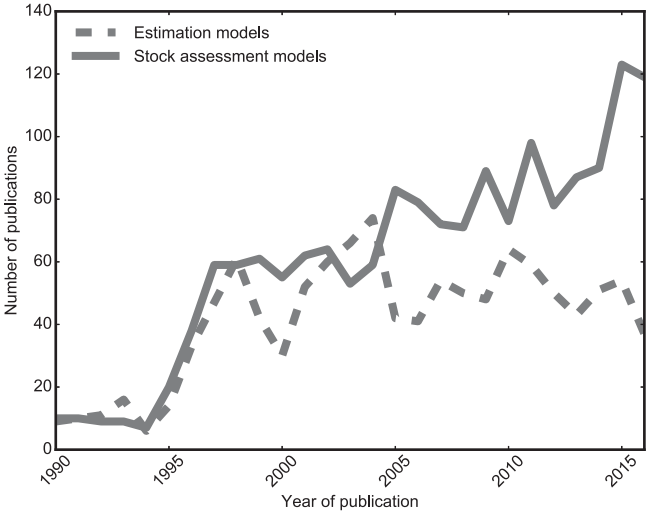


Figure 2. The number of publications per year for publications related to the topic estimation model and stock assessment model.

publications, whereas in practice (i.e., in fisheries stock assessments for management), these two topics are connected and combined into one model but reflect the different phases of the model development (Hoggarth et al., 2006). The distribution of publication frequencies for both general modeling topics is shown in Figure 2, which highlights the increased research interest in stock assessments models compared to estimation models. Additionally, the top five publications with the highest topic prevalence for each of the two modeling topics, indicating to what extent the content of a publication relates to the modeling topic, are shown in Table 2.

Interestingly, only the topics of estimation models and stock assessment models were uncovered (both of which focus on the ecological dimension of fisheries), whereas topics on economic and social fisheries aspects were not found within the modeling publications. This finding might be a result of the selection of journals used in this study. Most of the included fisheries journals declare a multi-disciplinary or interdisciplinary scope, while some specifically include socioeconomic considerations and the human dimension as subjects of interest. Therefore, at least one social or economic modeling topic could be expected to be identified by the LDA model. Another reason for the absence of other modeling topics may be that fisheries are still perceived as a natural science. The ICES only recently established the Strategic Initiative on the Human Dimension (SIHD) "to support the integration of social and economic science into ICES work" (ICES, 2017), and the majority of the ICES workgroups still lack social science input (ICES, 2016). As a result, social scientists and economists may pursue publication of their models not in a journal related to fisheries, but rather in a journal related to their respective disciplines or having a broader scope, such as *Ecology and Society, Marine Resource Economics* or *Marine Policy*. Merit issues could also contribute to the topic bias. Different scientific disciplines receive publication merits for different journals, which is more often dependent on the index of a journal (e.g., Science Citation Index (SCI), Social Science Citation Index (SSCI), or International Scientific Index (ISI)) than on its impact factor. As a result, non-biological and non-ecological disciplines are less likely to use top-tier fisheries journals as publication outlets. This might, in turn, lead to low visibility of non-ecological models among fisheries stakeholders, because many fisheries journals such as *Fish and Fisheries* and *Fisheries Research* intend to reach fisheries managers, administrators, policy makers, and legislators.

Table 2. Publication title, year, and topic prevalence (in percentages) for the five publications with the highest topic prevalence for each general modeling topic.

Modeling Topic	Title	Year	Prevalence
Estimation models	- Trawl survey based abundance estimation using datasets with unusually large catches.	1999	95.69%
	- Covariances in multiplicative estimates.	1999	94.35%
	- Use of simulation–extrapolation estimation in catch–effort analyses.	1999	93.90%
	- Reducing bias and filling in spatial gaps in fishery dependent catch per unit effort data by geostatistical prediction I methodology and simulation.	2014	92.23%
	- Confidence intervals for trawlable abundance from stratified-random bottom trawl surveys.	2011	90.48%
Stock assessment models	- The structure of complex biological reference points and the theory of replacement.	2009	99.37%
	- Analytical models for fishery reference points.	1998	98.50%
	- Implications of life-history invariants for biological reference points used in fishery management.	2003	98.14%
	- The estimation and robustness of FMSY and alternative fishing mortality reference points associated with high long-term yield.	2012	97.33%
	- Age-specific natural mortality rates in stock assessments: size-based vs. density-dependent.	2014	94.87%

(1) CATCH AND ABUNDANCE

word	prob.
MODELS	.013
CATCH	.011
ABUNDANCE	.008
SPECIES	.007
YEAR	.006
DEPTH	.006
EFFECTS	.005
CPUE	.005
VARIABLES	.005
SPATIAL	.004
LONGLINE	.004
LINEAR	.004
ENVIRONMENTAL	.004
EFFECT	.004
RATES	.004

(2) MORTALITY RATE (TAGS)

word	prob.
TAG	.016
MORTALITY	.014
RATES	.013
TAGGING	.013
RATE	.012
TAGS	.009
TAGGED	.009
MOVEMENT	.008
REPORTING	.006
MODELS	.006
YEAR	.006
FISHING	.006
RELEASE	.006
PARAMETERS	.005
FISHERY	.005

(3) ABUNDANCE (SURVEYS)

word	prob.
SPATIAL	.015
SURVEY	.011
ABUNDANCE	.009
DENSITY	.009
AREA	.009
ACOUSTIC	.007
VARIANCE	.007
SURVEYS	.006
SAMPLING	.006
DISTANCE	.005
BIOMASS	.005
RANDOM	.005
ESTIMATION	.004
SEA	.004
KM	.004

(4) RECREATIONAL FISHERIES

word	prob.
CATCH	.023
EFFORT	.015
FISHING	.012
SAMPLING	.012
SURVEY	.010
ANGLERS	.008
HARVEST	.007
SURVEYS	.007
RATE	.007
ANGLER	.007
FISHERY	.006
RECREATIONAL	.006
DAY	.005
VARIANCE	.005
LAKE	.005

(5) PARAMETERS AND ESTIMATORS

word	prob.
ERROR	.011
ABUNDANCE	.010
YEAR	.009
STOCK	.007
VARIANCE	.007
CATCH	.007
POPULATION	.006
MODELS	.006
INDEX	.006
YEARS	.005
ERRORS	.005
BIAS	.005
INDICES	.005
SAMPLE	.004
REGRESSION	.004

(6) SAMPLING

word	prob.
SAMPLING	.011
FISHING	.010
SPECIES	.010
FISHERY	.009
BYCATCH	.008
CATCH	.008
TRIP	.006
TRIPS	.006
OBSERVER	.006
VESSELS	.006
EFFORT	.005
SHRIMP	.005
LANDINGS	.005
VESSEL	.004
COMMERCIAL	.004

(7) ABUNDANCE (SAMPLING)

word	prob.
SAMPLING	.009
ABUNDANCE	.008
POPULATION	.007
BAYESIAN	.007
POSTERIOR	.007
PROBABILITY	.006
SPECIES	.006
CATCHABILITY	.006
MODELS	.006
CAPTURE	.006
DENSITY	.006
PRIOR	.005
SITES	.004
PARAMETERS	.004
ELECTROFISHING	.004

(8) FISH DISTRIBUTION

word	prob.
CATCH	.015
FISHING	.014
EFFORT	.013
FISHERY	.013
CPUE	.011
AREA	.011
COD	.011
ABUNDANCE	.010
CATCHABILITY	.009
BIOMASS	.008
STOCK	.006
AREAS	.006
SEASON	.006
CRAB	.006
RATES	.006

(9) SPAWNING

word	prob.
SPAWNING	.017
EGG	.014
EGGS	.012
PRODUCTION	.008
DAY	.007
STAGE	.007
BIOMASS	.006
LARVAE	.006
SAMPLING	.005
MORTALITY	.005
DAILY	.005
SAMPLES	.005
LARVAL	.005
TEMPERATURE	.004
FEMALES	.004

(10) NET SELECTIVITY

word	prob.
SELECTIVITY	.026
MESH	.013
LENGTH	.012
NET	.010
GILLNET	.009
SELECTION	.009
CATCH	.008
GEAR	.008
CURVE	.008
NETS	.007
CURVES	.007
GILL	.006
PARAMETERS	.006
MM	.006
RELATIVE	.006

(11) VESSELS AND FLEET

word	prob.
FISHING	.026
CATCH	.016
VESSEL	.012
EFFORT	.010
VESSELS	.010
FISHERY	.008
FLEET	.006
SPECIES	.006
CPUE	.006
POWER	.005
AREA	.004
YEAR	.004
MODELS	.004
RATE	.004
INFORMATION	.003

(12) TRAWL SURVEYS

word	prob.
SURVEY	.021
TRAWL	.019
SAMPLING	.013
SPECIES	.011
SURVEYS	.008
BOTTOM	.007
SAMPLE	.006
TOW	.006
LENGTH	.006
EFFICIENCY	.005
DESIGN	.005
AREA	.005
CATCH	.005
DENSITY	.005
TOWS	.005

(13) LENGTH AND GROWTH

word	prob.
LENGTH	.015
GROWTH	.014
PARAMETERS	.010
SAMPLE	.008
PARAMETER	.006
SAMPLES	.006
LIKELIHOOD	.006
ERROR	.005
MODELS	.005
STOCK	.005
FUNCTION	.005
DISTRIBUTIONS	.004
ESTIMATION	.004
STANDARD	.004
SET	.003

(14) SALMON

word	prob.
SALMON	.016
RIVER	.009
COUNTS	.007
SAMPLING	.007
ABUNDANCE	.007
RUN	.006
SURVEY	.005
SPAWNING	.004
POPULATION	.004
YEARS	.004
CHINOOK	.004
COUNT	.004
SAMPLE	.004
STREAM	.004
ESTIMATOR	.004

Figure 3. The 14 uncovered subtopics from the documents ($N = 1124$) exhibiting the topic estimation models as the dominant topic. The figure displays the subtopic label (top) and the top 15 high-probability words.

3.2. Subtopics within estimation models

The zoom-in (i.e., the process of uncovering subtopics from general topics) on the general topic of estimation models ($N = 1124$ documents) identified 14 subtopics (see Appendix 3). Figure 3 provides an overview of the 14 estimation model subtopics, the top 15 words of the topics with their probabilities, and the manually attached label that best captures the semantics of the top words. Furthermore, a two-dimensional topic representation can be found in the topic similarity map in Figure 4A, showing the topic similarity with respect to the distribution of the words. The trends (i.e., the change in overall topic proportion, in percentage points) and prevalence (i.e., the size of the overall topic proportion as a percentage) are presented in Figure 5A.

Most of the uncovered subtopics can be grouped. The principal group consists of the five subtopics focusing on the biological aspects of fisheries (i.e., catch and abundance, mortality rate (tags), fish distribution, spawning, and length and growth). This highlights the importance and scientific focus of the biological dimension in fisheries research. Catch and abundance shows the biggest overall increase over time (+15.46%) and had the largest proportion (14.84%) within the last six years (Figure 5A). Most of the other biological subtopics show very little variation over time, and some

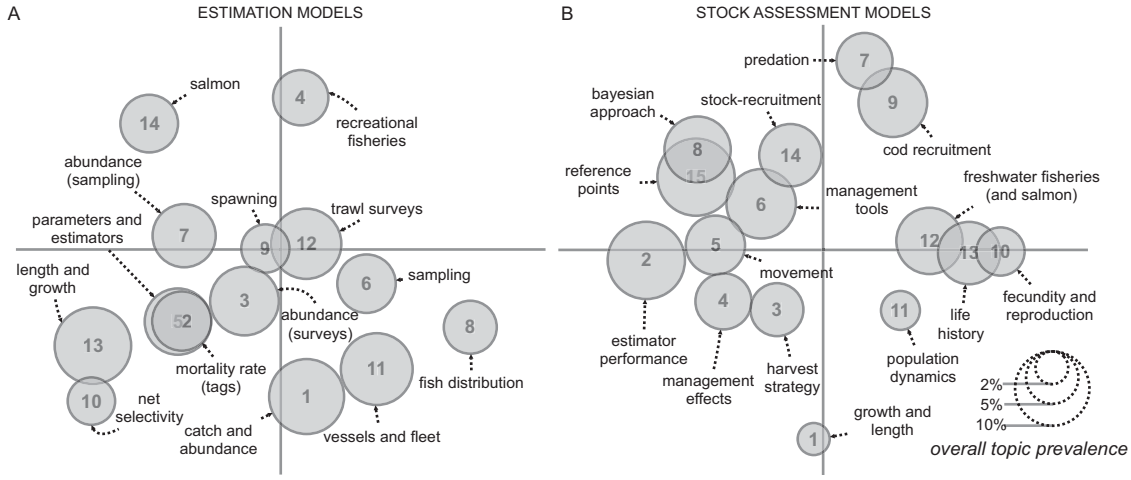


Figure 4. Topic similarity map that shows a two-dimensional representation (via multi-dimensional scaling). **A**: 14 estimation model subtopics. **B**: 15 stock assessment model subtopics. The distance between the nodes represents the topic similarity with respect to the distributions of the words (i.e., nodes closer together have more related word probabilities). The surface of the nodes represents the prevalence of the topic within the corpus.

only make a small contribution in terms of proportion (e.g., spawning), with only 3.82% overall topic proportion (Figure 5A). Length and growth showed the highest overall decrease over time (−14.04%), indicating a diminishing scientific interest. The subtopic of length and growth remained relatively high in terms of topic proportion, with an average of 9.13% between 2010 and 2016, possibly because growth is an important parameter for stock assessments (Lorenzen, 2016; Maunder et al., 2016) and is also most frequently discussed in fisheries, as shown by a previous trend analysis

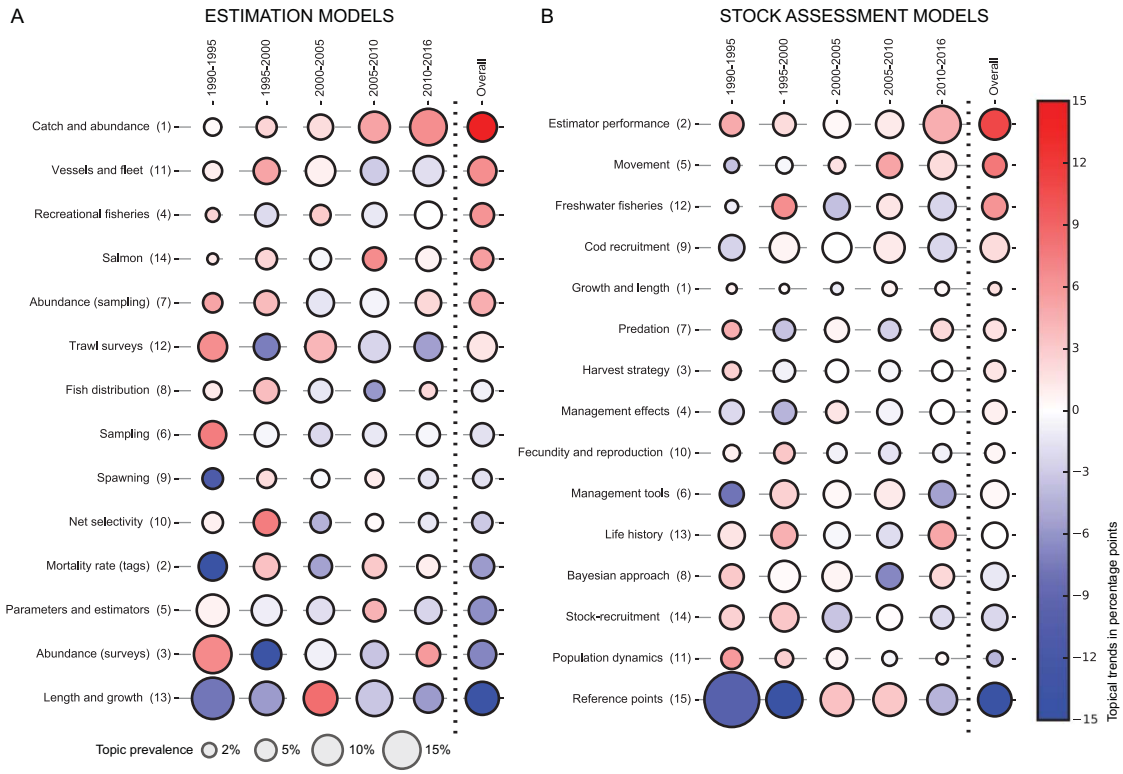


Figure 5. Trends in changing topic proportions for different time intervals for all subtopics. The left-hand side (**A**) displays the 14 uncovered estimation model subtopics. The surface of the node represents the topic prevalence within a certain time range and indicates how present a topic was within all the published material of that time frame. The colors indicate the trend in topic proportion (i.e., change in percentage points) and indicate whether a topic increased in popularity (hot topic) or decreased in popularity (cold topic) within that time frame. The right-hand side (**B**) displays the information for the 15 uncovered stock assessment model subtopics.

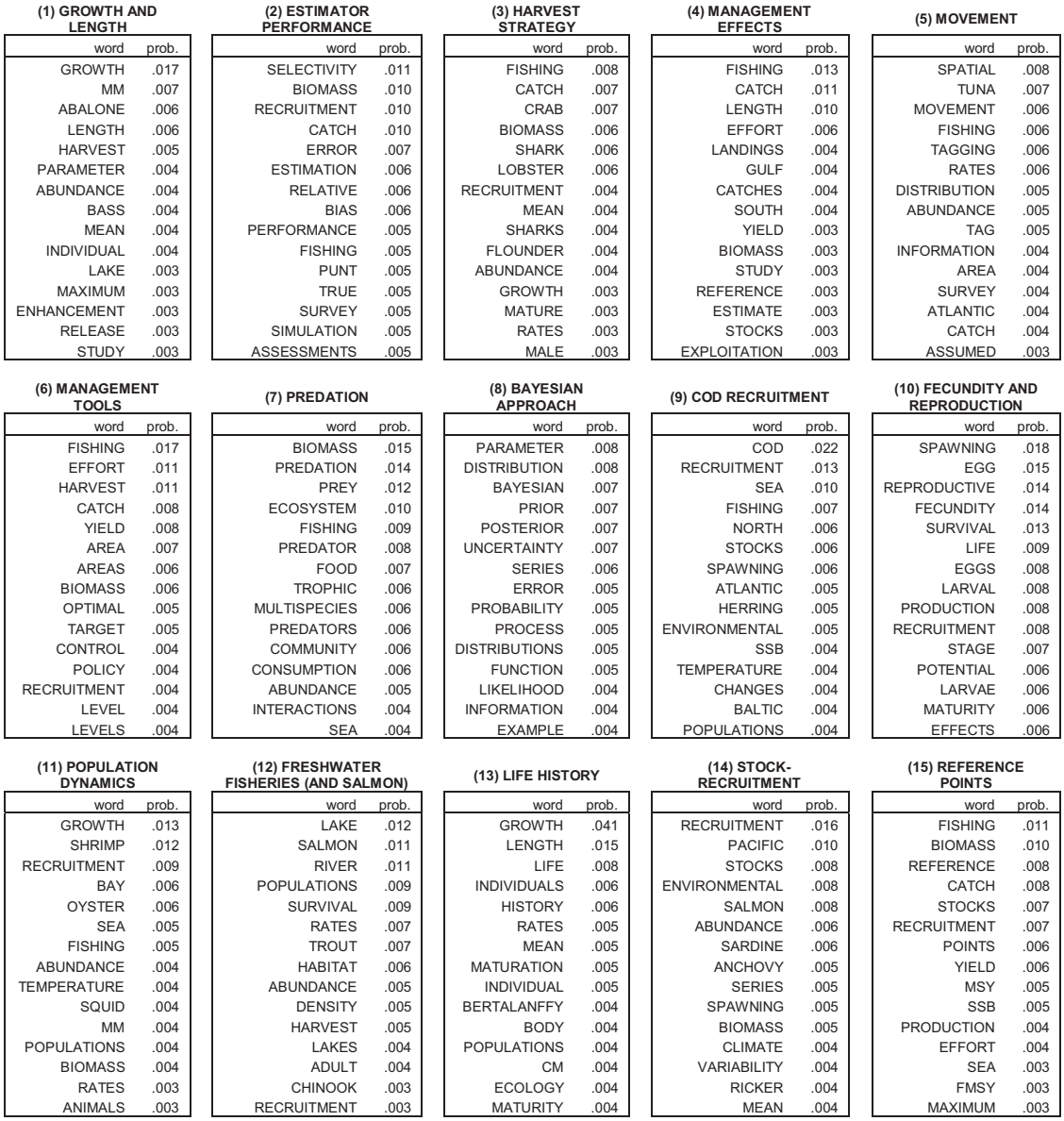
(Jarić et al., 2012). The subtopic of parameters and estimators relates more to the technical aspects of estimation modeling, but appears to be similar to the biological subtopic of mortality rate, as apparent from the similarity map (Figure 4A). Vessel and fleet showed a large topic proportion (between 8% and 10%) over the last 16 years (Figure 5A). Both the topic of vessel and fleet and that of net selectivity likely relate to biological considerations, but they could also hint at a slightly more economic perspective on industry (fleet) and gear-related matters; however, additional words such as "firm," "prices," or "market" would have to be present to confirm this hypothesis further. The four subtopics of abundance (survey), sampling, abundance (sampling), and trawl survey focus on survey and sampling, which are essential methods for gathering data and information on fisheries. In particular, information on catch and stock abundance is required by almost all stock assessment models (Hoggarth et al., 2006). These four subtopics account for a combined overall topic prevalence of 30.73%, indicating their importance to fisheries research. The subtopic of recreational fisheries refers to a type of fishery that differs in the estimation process compared to commercial fisheries, as it often employs surveys on anglers. This type of estimation process may refer not only to marine but also to freshwater fisheries. Recreational fisheries underwent an increase in topic proportion from 2.11% in the 1990–1995 period to 7.90% in the 2010–2016 period, indicating the growing importance of recreational fisheries assessments in fisheries science. The increased importance of recreational fishing on the commercial fish stocks (Griffiths and Fay, 2015) is in line with the observed trend in this study. Apart from recreational fisheries, no other types of fisheries (e.g., small-scale, artisanal, or commercial fisheries) were identified by the topic model. The distance of recreational fisheries from the other subtopics in the similarity map may explain this, as authors writing about recreational fisheries use distinctive words that are different from the discourse on other types of fisheries. Another possible explanation may be that there are more studies on recreational fisheries than on other types of fisheries. Salmon is the only topic that focuses on one particular species. The similarity map shows how the topic of salmon differs within the words used, indicating the particularity and specialized research niche of the topic (Figure 4A). Salmon showed a positive trend (+5.61%) over the study period; however, this result is in conflict with previous research that showed a diminishing research interest in the species (Jarić et al., 2012). This could be due to the increasing effort within aquaculture and the growing economic importance of the species over the period (FAO, 2016) that separates this study from that of Jarić, Cvijanović, Knežević-Jarić, and Lenhardt (2012).

Within the top 15 words of the subtopics, important subjects such as species and names/methods can be identified. Three subtopics contain species names (i.e., "shrimp" in sampling, "cod" and "crab" in fish distribution, and "salmon" and "chinook" in salmon). Methods mentioned within the subtopics of estimation models are "regression" in parameters and estimators and "Bayesian" in abundance (sampling). Parameters for fish stock assessments can be estimated through the least square method, represented in the form of regression analysis; however, maximum likelihood methods are now preferred, as they allow for a better specification in the form of errors in the models. Bayesian methods are commonly used to incorporate uncertainty into management advice, but this could also involve other methods such as maximum likelihood, bootstrapping, or Monte-Carlo modeling (Hoggarth et al., 2006). The two methods "regression" and "Bayesian" do not reflect the current diversity of modeling methods, nor necessarily the most conventional models used in fisheries assessments today, but they seem to have a strong association with the two topics of parameters and estimators and abundance (sampling). Note that references to names of species and methods highlight the importance and relation of such words within a specific topic – technically, they co-occur more frequently to describe the latent topic – but are by no means mutually exclusive (i.e., methods and species can occur in different subtopics simultaneously). They provide information from a topical perspective (i.e., a high-level decomposition of the document into clusters of co-occurring words), but fail to address on what basis such species and methods are linked within a specific topic.

3.3. Subtopics within stock assessment models

The zoom-in on the topic of stock assessment models ($N = 1637$ documents) revealed 15 subtopics (see Appendix 3 for the calculated topic coherence scores). Figure 6 provides an overview of the 15 subtopics, the top 15 words with their probabilities, and the label attached to each topic. The topic similarity for these subtopics can be found in Figure 4B. The subtopic trends and prevalence are displayed in Figure 5B.

Most of the subtopics of stock assessment models evolve around biological aspects and processes (i.e., growth and length, movement, predation, cod recruitment, fecundity and reproduction, population dynamics, life history, and stock recruitment). The majority of these subtopics show a slight increase over the study period (Figure 5B); together, these subtopics have an overall topic proportion of 42.91%, which shows their consistent importance within fisheries science and fisheries management (Hilborn and Walters, 1992). Within the biological subtopics, predation stands out as the only subtopic that refers to "interaction," "multi-species," and the "ecosystem." The subtopic of predation increased by 4.67% during the period from 1990 to 1995



(1) GROWTH AND LENGTH

word	prob.
GROWTH	.017
MM	.007
ABALONE	.006
LENGTH	.006
HARVEST	.005
PARAMETER	.004
ABUNDANCE	.004
BASS	.004
MEAN	.004
INDIVIDUAL	.004
LAKE	.003
MAXIMUM	.003
ENHANCEMENT	.003
RELEASE	.003
STUDY	.003

(2) ESTIMATOR PERFORMANCE

word	prob.
SELECTIVITY	.011
BIOMASS	.010
RECRUITMENT	.010
CATCH	.010
ERROR	.007
ESTIMATION	.006
RELATIVE	.006
BIAS	.006
PERFORMANCE	.005
FISHING	.005
PUNT	.005
TRUE	.005
SURVEY	.005
SIMULATION	.005
ASSESSMENTS	.005

(3) HARVEST STRATEGY

word	prob.
FISHING	.008
CATCH	.007
CRAB	.007
BIOMASS	.006
SHARK	.006
LOBSTER	.006
RECRUITMENT	.004
MEAN	.004
SHARKS	.004
FLOUNDER	.004
ABUNDANCE	.004
GROWTH	.003
MATURE	.003
RATES	.003
MALE	.003

(4) MANAGEMENT EFFECTS

word	prob.
FISHING	.013
CATCH	.011
LENGTH	.010
EFFORT	.006
LANDINGS	.004
GULF	.004
CATCHES	.004
SOUTH	.004
YIELD	.003
BIOMASS	.003
STUDY	.003
REFERENCE	.003
ESTIMATE	.003
STOCKS	.003
EXPLOITATION	.003

(5) MOVEMENT

word	prob.
SPATIAL	.008
TUNA	.007
MOVEMENT	.006
FISHING	.006
TAGGING	.006
RATES	.006
DISTRIBUTION	.005
ABUNDANCE	.005
TAG	.005
INFORMATION	.004
AREA	.004
SURVEY	.004
ATLANTIC	.004
CATCH	.004
ASSUMED	.003

(6) MANAGEMENT TOOLS

word	prob.
FISHING	.017
EFFORT	.011
HARVEST	.011
CATCH	.008
YIELD	.008
AREA	.007
AREAS	.006
BIOMASS	.006
OPTIMAL	.005
TARGET	.005
CONTROL	.004
POLICY	.004
RECRUITMENT	.004
LEVEL	.004
LEVELS	.004

(7) PREDATION

word	prob.
BIOMASS	.015
PREDATION	.014
PREY	.012
ECOSYSTEM	.010
FISHING	.009
PREDATOR	.008
FOOD	.007
TROPHIC	.006
MULTISPECIES	.006
PREDATORS	.006
COMMUNITY	.006
CONSUMPTION	.006
ABUNDANCE	.005
INTERACTIONS	.004
SEA	.004

(8) BAYESIAN APPROACH

word	prob.
PARAMETER	.008
DISTRIBUTION	.008
BAYESIAN	.007
PRIOR	.007
POSTERIOR	.007
UNCERTAINTY	.007
SERIES	.006
ERROR	.005
PROBABILITY	.005
PROCESS	.005
DISTRIBUTIONS	.005
FUNCTION	.005
LIKELIHOOD	.004
INFORMATION	.004
EXAMPLE	.004

(9) COD RECRUITMENT

word	prob.
COD	.022
RECRUITMENT	.013
SEA	.010
FISHING	.007
NORTH	.006
STOCKS	.006
SPAWNING	.006
ATLANTIC	.005
HERRING	.005
ENVIRONMENTAL	.005
SSB	.004
TEMPERATURE	.004
CHANGES	.004
BALTIC	.004
POPULATIONS	.004

(10) FECUNDITY AND REPRODUCTION

word	prob.
SPAWNING	.018
EGG	.015
REPRODUCTIVE	.014
FECUNDITY	.014
SURVIVAL	.013
LIFE	.009
EGGS	.008
LARVAL	.008
PRODUCTION	.008
RECRUITMENT	.008
STAGE	.007
POTENTIAL	.006
LARVAE	.006
MATURITY	.006
EFFECTS	.006

(11) POPULATION DYNAMICS

word	prob.
GROWTH	.013
SHRIMP	.012
RECRUITMENT	.009
BAY	.006
OYSTER	.006
SEA	.005
FISHING	.005
ABUNDANCE	.004
TEMPERATURE	.004
SQUID	.004
MM	.004
POPULATIONS	.004
BIOMASS	.004
RATES	.003
ANIMALS	.003

(12) FRESHWATER FISHERIES (AND SALMON)

word	prob.
LAKE	.012
SALMON	.011
RIVER	.011
POPULATIONS	.009
SURVIVAL	.009
RATES	.007
TROUT	.007
HABITAT	.006
ABUNDANCE	.005
DENSITY	.005
HARVEST	.005
LAKES	.004
ADULT	.004
CHINOOK	.003
RECRUITMENT	.003

(13) LIFE HISTORY

word	prob.
GROWTH	.041
LENGTH	.015
LIFE	.008
INDIVIDUALS	.006
HISTORY	.006
RATES	.005
MEAN	.005
MATURATION	.005
INDIVIDUAL	.005
BERTALANFFY	.004
BODY	.004
POPULATIONS	.004
CM	.004
ECOLOGY	.004
MATURITY	.004

(14) STOCK-RECRUITMENT

word	prob.
RECRUITMENT	.016
PACIFIC	.010
STOCKS	.008
ENVIRONMENTAL	.008
SALMON	.008
ABUNDANCE	.006
SARDINE	.006
ANCHOVY	.005
SERIES	.005
SPAWNING	.005
BIOMASS	.005
CLIMATE	.004
VARIABILITY	.004
RICKER	.004
MEAN	.004

(15) REFERENCE POINTS

word	prob.
FISHING	.011
BIOMASS	.010
REFERENCE	.008
CATCH	.008
STOCKS	.007
RECRUITMENT	.007
POINTS	.006
YIELD	.006
MSY	.005
SSB	.005
PRODUCTION	.004
EFFORT	.004
SEA	.003
FMSY	.003
MAXIMUM	.003

Figure 6. The 15 uncovered subtopics from the documents ($N = 1637$) exhibiting the topic stock assessment models as the dominant topic. The figure displays the subtopic label (top) and the top 15 high-probability words.

(Figure 5B), which reflects the increased scientific awareness of predator–prey interaction and model implications in the early 1990s (e.g., Yodzis, 1994). The topic proportion of predation shows a positive trend, as it rose from 3.75% in the period of 1990–1995 to 5.07% in the period of 2010–2016; this might indicate the increased attention of the scientific community to an ecosystem approach to fisheries and the implementation of multi-species and ecosystem considerations within stock assessments, modeling frameworks, and management advice (Maynou, 2014; Möllmann et al., 2014; Gaichas et al*i*, 2017). The four subtopics of harvest strategy, management effects, management tools and reference points all concern management measures and effects, but they mainly address biological components such as "recruitment," "abundance," and "biomass." Reference points shows the strongest overall negative trend of all subtopics (−26.55%), indicating that the popularity of this topic among fisheries scientists has decreased over the years. Nevertheless, the topic of reference points still makes up a relatively large proportion, 9.82% (Figure 5B); this is the second largest proportion in the period of 2010–2016 after estimator performance, which has a 15.19% topic proportion within the same period. This highlights the continuity of research on reference points from the 1990s to the present day (Caddy and Mahon, 1995; Caddy, 2004; Froese et al., 2017). The subtopic of estimator performance shows the highest increase (+11.11%) within the overall study period (i.e., 1990–2016) and makes up a large proportion within the last six years of the time frame, from 2010 to 2016 (15.19%); this finding could be related to the increased

overall importance of models in fisheries science (Jarić et al., 2012). The subtopic of freshwater fisheries shows an overall positive trend (+6.28%), even though freshwater fisheries habitats have been found to be less studied than marine fisheries (Jarić et al., 2012). The topic proportion of freshwater fisheries rose over the study period, from 1.82% in 1990–2000 to 8.08% in 2010–2016 (Figure 5B). The importance of freshwater fisheries in areas such as Africa and India may explain the increase in research efforts within this field (FAO, 2016).

From the top 15 words (Figure 6), related subjects were identified, such as regions, species, and names/methods. The two marine regions mentioned are "Atlantic" and "Pacific," possibly because these are some of the world's major fishing areas (FAO, 2016). The various species names found within the top 15 words, such as "cod," "herring," and "anchovy," cover many of the commercially important species in marine capture production (FAO, 2016). These results stand in stark contrast to a bibliometric study on trends in fisheries science, which found virtually no research on many commercially important species (Aksnes and Browman, 2016); however, these results were based on word frequencies in publication titles and abstracts, which may not mention the species of concern. This finding highlights the strength of the full-text LDA analysis. Other mentioned species, such as "abalone," "lobster," and "shark," may have high probabilities for occurrence in the subtopics because they represent species of great economic value and also are often a focus of conservation efforts (Turpie et al., 2003; Simpfendorfer and Dulvy, 2017).

Several names within the words of the subtopics refer to a method named after a scientist, e.g., "Bayesian," "Bertalanffy," "Ricker," and "Punt," which could be a direct consequence of the inclusion of the reference list in the analysis. The subtopic of Bayesian approach indicates the importance of this methodology in fishery science and for fisheries models. A Bayesian approach can be used for stock assessments and decision analysis and resembles an improved way of fitting models to data and decision-making (Hoggarth et al., 2006). The scientists von Bertalanffy and Ricker both made substantial contributions to fisheries science – von Bertalanffy in metabolism and growth (von Bertalanffy, 1957) and Ricker in the computation and interpretation of computational statistics of fish populations (Ricker, 1975). Their methods are still applied today in the form of growth models (Allen, 1966; Piner et al., 2016) and in stock-recruitment models (Baker et al., 2014). The author Punt has not developed any particular method that takes his name; however, his name may occur within the top 15 words due to his significant contribution to research and his publications on estimator performance and data standardization, as well as his many citations by other scientists within the field. Although Punt is, relatively speaking, a newcomer compared to some of the early influential researchers in the field (e.g., Hjort, Beverton, and Holt), the occurrence of his name is perhaps a result of the timeframe examined, or it may indicate that the names of senior scientists and methods have become somewhat common knowledge and are therefore not always explicitly stated or cited.

4. Conclusions

The aim of this paper was to uncover fisheries modeling topics from 22,236 scientific publications from 13 peer-reviewed fisheries journals. Additionally, subtopics from general modeling topics were uncovered to provide insights into their developments and trends over the last 26 years. Overall, two main fisheries modeling topics were identified: estimation models and stock assessment models. This study demonstrates that research in the field of fisheries modeling shows a shift of scientific focus in topics and subtopics over the last 26 years. Stock assessment models are outperforming estimation models, and their underlying subtopics have moved from length and growth to catch and abundance, and from reference points to estimator performance over the last 26 years. Economically important species and areas show a high presence within the modeling subtopics.

Both general modeling topics focus primarily on the biological aspects of fisheries; however, since this study was limited to publications in 13 fisheries journals, other topics in fisheries modeling (e.g., with a focus on social, management or economic aspects of fisheries) may well exist in publications of other journals. Possible disciplinary merit issues and the remaining understanding of fisheries as a natural science discipline might further limit fisheries journals to models with an ecological focus, despite their multi-disciplinary scope.

In conclusion, this novel machine learning approach revealed interesting insights into the topical trends of a large dataset of models published in fisheries journals. This approach enables researchers to identify research topics and shifts in research focus, and it provides a bigger picture that captures the main ideas prevailing scientific publications.

Acknowledgments

We are grateful to John Pope, Melania Borit, and several anonymous reviewers for improving earlier versions of this article.

Funding

This research was funded by the project SAF21 – Social Science Aspects of Fisheries for the 21st Century (project financed under the EU Horizon 2020 Marie Skłodowska-Curie (MSC) ITN-ETN Program; project number: 642080).

ORCID

Shaheen Syed http://orcid.org/0000-0001-5462-874X
Charlotte Teresa Weber http://orcid.org/0000-0003-4371-695X

References

Aksnes, D. W., and H. I. Browman. An overview of global research effort in fisheries science. *ICES J. Mar. Sci.: J. du Conseil*, **73**(4): 1004–1011 (2016). doi:10.1093/icesjms/fsv248.

Aletras, N., and M. Stevenson. Evaluating topic coherence using distributional semantics. **In**: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Association for Computational Linguistics, 13–22 (2013).

Allen, K. R. A method of fitting growth curves of the von bertalanffy type to observed data. *J. Fish. Res. Board Can.*, **23**(2): 163–179 (1966). doi:10.1139/f66-016.

Angelini, R., and C. L. Moloney. Fisheries, ecology and modelling: An historical perspective. *Pan-Am. J. Aquat. Sci.*, **2**(2): 75–85 (2007).

Asuncion, A., M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. *Proc. Twenty-Fifth Conf. Uncertainty Artif. Intell.*, (Ml): 27–34 (2012).

Baker, M. R., D. E. Schindler, T. E. Essington, and R. Hilborn. Accounting for escape mortality in fisheries: Implications for stock productivity and optimal management. *Ecol. Appl.*, **24**(1): 55–70 (2014). doi:10.1890/12-1871.1.

Bastardie, F., J. R. Nielsen, and T. Miethe. DISPLACE: A dynamic, individual-based model for spatial fishing planning and effort displacement — integrating underlying fish population models. *Can. J. Fish Aquat. Sci.*, **71**(3): 366–386 (2014). doi:10.1139/cjfas-2013-0126.

von Bertalanffy, L. Quantitative laws in metabolism and growth. *Q. Rev. Biol.*, **32**(3): 217–231 (1957). doi:10.1086/401873.

Bjørndal, T., D. E. Lane, and A. Weintraub. Operational research models and the management of fisheries and aquaculture: A review. *Eur. J. Oper. Res.*, **156**(3): 533–540 (2004). doi:10.1016/S0377-2217(03)00107-3.

Blei, D. M. Communications of the ACM, **55**(4): 77–84 (2012). doi: 10.1145/2133806.2133826.

Blei, D. M., and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Anal.*, **1**(1): 121–143 (2006). doi:10.1214/06-BA104.

Blei, D. M., and J. D. Lafferty. Dynamic topic models. **In**: *Proceedings of the 23rd international conference on Machine learning – ICML '06*. New York, NY: ACM Press, 113–120 (2006). doi:10.1145/1143844.1143859

Blei, D. M., A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Machine Learn. Res.*, **3**: 993–1022 (2003).

Bouma, G. Normalized (Pointwise) mutual information in collocation extraction. **In**: *Proceedings of German Society for Computational Linguistics (GSCL 2009)*. 31–40 (2009).

Caddy, J. F. Current usage of fisheries indicators and reference points, and their potential application to management of fisheries for marine invertebrates. *Can. J. Fish. Aquat. Scie.*, **61**(8): 1307–1324 (2004). doi:10.1139/f04-132.

Caddy, J. F., and R. Mahon. *Reference points for fisheries management*, vol. 347. Food and Agriculture Organization (FAO) Fisheries Technical Paper. Rome: FAO (1995).

Cadrin, S. X., and M. Dickey-Collas. Stock assessment methods for sustainable fisheries. *ICES J. Mar. Sci.*, **72**(1): 1–6 (2015). doi:10.1093/icesjms/fsu228.

Chuang, J., D. Ramage, C. D. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. pp. 443–452. **In**: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (G. Rebecca, T. Rodden, P. Aoki, E. Cutrell, J. Robin and G. Olson, Eds). Montreal, Canada: ACM.

Debortoli, S., O. Müller, I. Junglas, and J. Vom Brocke. Text mining for information systems researchers: An annotated topic modeling tutorial. *Commun. Assoc. Inf. Syst.*, **39**: 110–135 (2016).

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, **41**(6): 391–407 (1990). doi:10.1002/(SICI)1097-4571(199009)41:6%3c391::AID-ASI1%3e3.0.CO;2-9.

DiMaggio, P., M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, **41**(6): 570–606 (2013). doi:10.1016/j.poetic.2013.08.004.

Epstein, J. M. Why model? *J. Artif. Soc. Soc. Simul.*, **11**(4): 12 (2008).

Evangelopoulos, N., X. Zhang, and V. R. Prybutok. Latent Semantic Analysis: Five methodological recommendations. *Eur. J. Inf. Syst.*, **21**(1): 70–86 (2012). doi:10.1057/ejis.2010.61.

FAO. The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all. Rome (2016).

Froese, R., N. Demirel, G. Coro, K. M. Kleisner, and H. Winker. Estimating fisheries reference points from catch and resilience. *Fish Fish.*, **18**(3): 506–526 (2017). doi:10.1111/faf.12190.

Gaichas, S. K., M. Fogarty, G. Fay, R. Gamble, S. Lucey, and L. Smith. Combining stock, multispecies, and ecosystem level fishery objectives within an operational management procedure: Simulations to start the conversation. *ICES J. Mar. Sci.: J. du Conseil*, **74**(2): 552–565 (2017).

Le Gallic, B. Fisheries Sustainability Indicators : The OECD experience. Joint workshop on "Tools for measuring (integrated) Fisheries Policy aiming at sustainable ecosystem". Brussels: OECD (2002).

Gerl, T., H. Kreibich, G. Franco, D. Marechal, and K. Schröter. A review of flood loss models as basis for harmonization and benchmarking. *PLOS ONE*, **11**(7): e0159791 (2016). doi:10.1371/journal.pone.0159791.

Griffiths, S. P., and G. Fay. Integrating recreational fisheries data into stock assessment: Implications for model performance and subsequent harvest strategies. *Fish. Manage. Ecol.*, **22**(3): 197–212 (2015). doi:10.1111/fme.12117.

Griffiths, T. L., and M. Steyvers. Finding scientific topics. *Proc. Nat. Acad. Sci.*, **101**(Supplement 1): 5228–5235 (2004). doi:10.1073/pnas.0307752101.

Griffiths, T. L., M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychol. Rev.*, **114**(2): 211–244 (2007).

Haapasaari, P., S. Mäntyniemi, and S. Kuikka. Baltic herring fisheries management: Stakeholder views to frame the problem. *Ecol. Soc.*, **17**(3): art36 (2012). doi:10.5751/ES-04907-170336.

Hilborn, R., and C. J. Walters Quantitative fisheries stock assessment : Choice, dynamics and uncertainty. New York, NY: Chapman & Hall (1992).

Hoffman, M. D., D. M. Blei, and F. Bach. Online learning for latent Dirichlet allocation. *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, Vol. 1, pp. 856–864. USA: Curran Associates Inc. (2010).

Hofmann, T. Probabilistic latent semantic indexing. **In**: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, USA, 50–57. New York: ACM. isbn = 1-58113-096-1 (1999). doi = 10.1145/312624.312649

Hoggarth, D. D., S. Abeyasekera, R. I. Arthur, J. R. Beddington, R. W. Burn, A. S. Halls, G. P. Kirkwood, M. McAllister, P. Medley, C. C. Mees, G. B. Parkes, G. M. Pilling, R. C. Wakeford, and R. L. Welcomme. Stock assessment for fishery management : A framework guide to the stock assessment tools of the fisheries management and science programme. *FAO Fish. Tech. Pap.*, **487** (2006).

ICES. Report of the SIHD survey of the current state of 'human dimension' in some ICES groups, 31 pp. (2016).

ICES. SIHD [online]. Strategic Initiative on the Human Dimension. Available from: http://www.ices.dk/community/groups/Pages/SIHD.aspx (2017).

Jarić, I., G. Cvijanović, J. Knežević-Jarić, and M. Lenhardt. Trends in fisheries science from 2000 to 2009: A bibliometric study. *Rev. Fish. Sci.*, **20**(2): 70–79 (2012). doi:10.1080/10641262.2012.659775.

Jennings, S. Indicators to support an ecosystem approach to fisheries. *Fish Fish.*, **6**(3): 212–232 (2005). doi:10.1111/j.1467-2979.2005.00189.x.

Larsen, P. O., and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, **84**(3): 575–603 (2010). doi:10.1007/s11192-010-0202-z.

Lorenzen, K. Toward a new paradigm for growth modeling in fisheries stock assessments: Embracing plasticity and its consequences. *Fish. Res.*, **180**: 4–22 (2016).

Maunder, M. N., P. R. Crone, A. E. Punt, J. L. Valero, and B. X. Semmens. Growth: Theory, estimation, and application in fishery stock assessment models. *Fish. Res.*, **180**: 1–3 (2016).

Maynou, F. Coviability analysis of western mediterranean fisheries under MSY scenarios for 2020. *ICES J. Mar. Sci.*, **71**(7): 1563–1571 (2014). doi:10.1093/icesjms/fsu061.

Minto, C., and C. Lordan. GEPETO: Review of mixed fisheries modelling approaches for the Celtic Sea (2014).

Möllmann, C., M. Lindegren, T. Blenckner, L. Bergström, M. Casini, R. Diekmann, J. Flinkman, B. Müller-Karulis, S. Neuenfeldt, J. O. Schmidt, M. Tomczak, R. Voss, and A. Gårdmark. Implementing ecosystem-based fisheries management: From single-species to integrated ecosystem assessment and advice for Baltic Sea fish stocks. *ICES J. Mar. Sci.*, **71**(5): 1187–1197 (2014). doi:10.1093/icesjms/fst123.

Oecd. Main science and technology indicators. *Sci. Technol.*, **2008**: 104 (2008).

Ostrom, E. A general framework for analyzing sustainability of social-ecological systems. *Science*, **325**(5939): 419–422 (2009). doi:10.1126/science.1172133.

Partelow, S. Key steps for operationalizing social–ecological system framework research in small-scale fisheries: A heuristic conceptual approach. *Mar. Pol.*, **51**: 507–511 (2015).

Piner, K. R., H.-H. Lee, and M. N. Maunder. Evaluation of using random-at-length observations and an equilibrium approximation of the population age structure in fitting the von Bertalanffy growth function. *Fish. Res.*, **180**(180): 128–137 (2016).

Plagányi, E. E. Models for an ecosystem approach to fisheries. FAO Fisheries Technical Paper, 477. Rome: FAO, pp. 10 (2007).

Prellezo, R., P. Accadia, J. L. Andersen, B. S. Andersen, E. Buisman, A. Little, J. R. Nielsen, J. J. Poos, J. Powell, and C. Röckmann. A review of EU bio-economic models for fisheries: The value of a diversity of models. *Mar. Pol.*, **36**(2): 423–431 (2012).

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. How to analyze political attention with minimal assumptions and costs. *Am. J. Polit. Sci.*, **54**(1): 209–228 (2010). doi:10.1111/j.1540-5907.2009.00427.x.

Rehurek, R., and P. Sojka. Software framework for topic modelling with large corpora, pp. 46–50. **In**: *Proceedings of the LREC 2010 Workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta. DOI: 10.13140/2.1.2393.1847. ISBN 2-9517408-6-7 (2010).

Ricker, W. E. Computation and interpretation of biological statistics of fish populations. *Bull. Fish. Res. Board Can.*, **191**: 1–382 (1975).

Röder, M., A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. **In**: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai, China, pp. 399–408. New York, NY, USA: ACM. ISBN 978-1-4503-3317-7 (2015).

Sievert, C., and K. Shirley. LDAvis: A method for visualizing and interpreting topics. **In**: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 63–70 (2014).

Simpfendorfer, C. A. and N. K. Dulvy. Bright spots of sustainable shark fishing. *Curr. Biol.*, **27**(3): R97–R98 (2017).

Srivastava, A., and M. Sahami. Text mining: Classification, clustering, and applications. Boca Raton, FL: CRC Press (2009).

Stevens, K., P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring topic coherence over many models and many topics. **In**: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 952–961 (2012).

Syed, S., and M. Spruit. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. **In**: *The 4th IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 165–174 (2017). doi:10.1109/DSAA.2017.61.

Teh, Y. W., D. Newman, M. Welling, and D. Neaman. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. **In**: *NIPS'06 Proceedings of the 19th International Conference on Neural Information Processing*

Systems, Canada. 1353–1360 MA, USA: MIT Press Cambridge (2006).

Turpie, J. K., B. J. Heydenrych, and S. J. Lamberth. Economic value of terrestrial and marine biodiversity in the cape floristic region: Implications for defining effective and socially optimal conservation strategies. *Biol. Conserv.*, **112**(1–2): 233–251 (2003).

Urquhart, C. An encounter with grounded theory : Tackling the practical and philosophical issues, pp. 104–140. *Qual. Res. Inf. Syst.: Issues Trends*, (E. M. Trauth, Ed.). Hershey, PA, USA: IGI Global (2001). DOI: 10.4018/978-1-930708-06-8.ch005.

Wang, C., J. Paisley, and D. M. Blei. Online variational inference for the hierarchical dirichlet process, vol 15, pp. 752–760. **In**: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, (G. Gordon, D. Dunson and M. Dudík, Eds.). *Proceedings of Machine Learning Research*. Fort Lauderdale, FL, USA: PMLR (2011).

Yodzis, P. Predator-prey theory and management of multispecies fisheries. *Ecol. Appl.*, **4**(1): 51–58 (1994).

Appendix 1

Generative Process of LDA

The generative process of LDA is described below:

1. For each topic
   a. Draw a distribution over all the words, $\beta_K \sim \mathrm{Dir}(\eta)$
2. For each document
   a. Draw a distribution over topics $\theta_d \sim \mathrm{Dir}(\alpha)$ (per-document topic proportion)
   b. For each word in the document
      i. Draw a topic $z_{d,n}$ from θ_d (per-word topic assignment)
      ii. Draw a word $w_{d,n}$ from that topic

Each topic is a multi-nomial distribution over all the words and arises from a Dirichlet distribution β_k $\mathrm{Dir}(\eta)$. Additionally, each document is represented as a distribution over the topics and arises from a Dirichlet distribution $\theta_\mathrm{d} \sim \mathrm{Dir}(\alpha)$. The Dirichlet parameter η defines the smoothing of the words within topics, and α defines the smoothing of the topics within documents. The per-word topic assignment $z_{d,n}$ is the topic drawn from the per-document topic proportions (Step 2a) for the n-th word in the d-th document. The joint distribution of the observed words w_D and the hidden variables β_K (topics), θ_D (document topic proportions), and z_D (word topic assignments) becomes:

$$p\ (\beta_K, \theta_D, z_D, w_D) = \prod_{k=1}^{K} p(\beta_K \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d)\ p\left(w_{d,n} \mid z_{d,n}, \beta_{d,k}\right) \quad (1)$$

The per-word topic assignment $z_{d,n}$ depends on the per-document topic proportion θ_d it draws a topic for each word from the previously drawn per-document topic proportion. As a result, the generative process creates documents that contain multiple topics in varying proportions. The drawn word $w_{d,n}$ depends on the per-word topic assignment $z_{d,n}$ (it draws a word from the previously drawn topic) and all the topics β_K (the probability of $w_{d,n}$ (row) is retrieved from $z_{d,n}$ (column) within the $K \times V$ topic matrix).

Equation 1 shows the joint probability of all the hidden and observed variables and the encoded statistical assumptions underlying LDA. The process now is to infer the hidden variables from the observed variables in order to obtain the topics and topic proportions per document. The inference is based on the conditional probability of the hidden variables given the observed words, also known as the posterior distribution (see Equation 2). Moreover, this inference can be viewed as a reversal of the generative process, and it tries to identify the structure likely to have generated the data.

$$p(\beta_K, \theta_D, z_D \mid w_D) = \frac{p(\beta_K, \theta_D, z_D, w_D)}{p(w_D)} \quad (2)$$

Unfortunately, the posterior is intractable to compute (Blei et al., 2003) due to the denominator. The marginal probability $p(w_D)$ is the sum of the joint distribution over all instances of the hidden structure and is exponentially large (Blei, 2012). The computational problem now is to estimate the posterior distribution using statistical inference techniques. Several methods exist, such as variational and sampling-based algorithms, for achieving a sufficiently close approximation of the true posterior (Blei and Jordan, 2006; Teh et al., 2006; Hoffman et al., 2010; Wang et al., 2011). Variational methods place a family of probability distributions onto the latent structure and aim to find the distribution closest to the true posterior, measured with, for example, Kullback–Leibler (KL) divergence. Sampling-based inference is a repeated sampling process, generally using one variable at a time while fixing the other variables, until the process converges; the sample values will have the same distribution as if they came from the true posterior. An example of sampling-based inference is the Gibbs sampler (Griffiths and Steyvers, 2004), a Markov chain Monte Carlo (MCMC) algorithm. It is important to note that both variational and sampling-based approaches provide similarly accurate results (Asuncion et al., 2012).

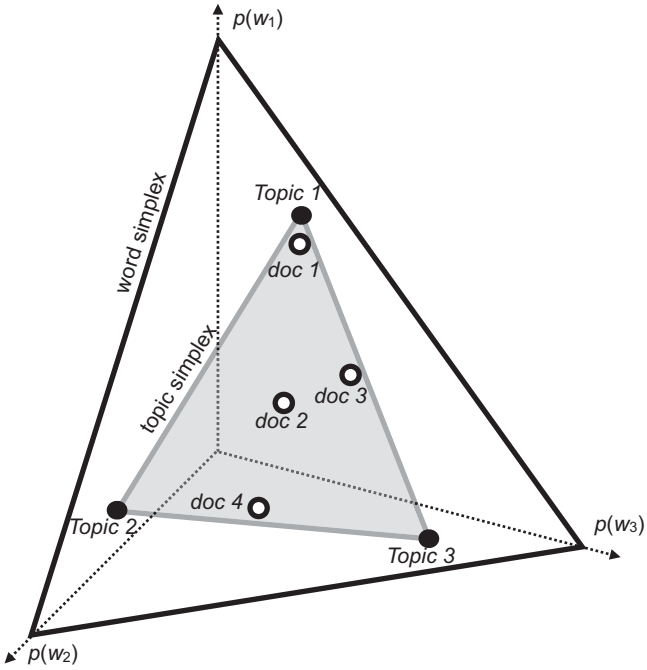


Figure 7. Geometric interpretation of LDA, showing a (V-1)-dimensional word simplex with $V = \{w1, w2, w3\}$, in which every point on the simplex represents a discrete distribution of word probabilities. A point closer to one of the corners indicates that more probability mass is placed on that word. Similarly, within the word simplex a topic simplex can be found, in which a topic represents some probability distribution over all the words (three words in this example). The documents, represented as distributions over topics, are placed within the (K-1)-dimensional topic simplex. As such, each document is represented as a discrete probability distribution over K topics, which in this example is three.

Figure 7 displays a simplified geometric interpretation of LDA. The vocabulary V contains just three words (w_1, w_2, w_3) and is represented as a (V-1)-dimensional word simplex. In reality, the word simplex contains many dimensions, as the vocabulary can easily contain thousands of words. The word simplex relates to all the probability distribution of words. Similarly, Figure 7 illustrates how the topics, modeled as distributions of the vocabulary, are positioned within the word simplex. The example shows three topics T, represented as a (T-1)-dimensional topic simplex. The documents, modeled as distributions over the topics, are points on the topic simplex. For example, Document 1 deals almost entirely with Topic 1; Document 2 exhibits all three topics in equal proportions; and Document 3 has equal proportions of Topics 1 and 3 but none of Topic 2. Note that this only holds if the topic simplex is defined by a uniform Dirichlet distribution that assigns equal probability mass to all topics. The shapes of the Dirichlet distributions within the word simplex and topic simplex are given by η and α, respectively.

Appendix 2

See Table 3

Table 3. The words that occurred in ≥90% of the documents and that are thus eliminated from the study. Words that occur in almost every document have no significant topical distinctiveness, and including them would cause these words to dominate every topic. N is the number of publications.

Dataset	N	Words
Overall	22,236	of, and, for, to, the, in, with, is, from, as, this, that, on, are, at, be, an, or, not, was, have, these, were, which, also, between, been, than, all, other, it, more, has, their, but, two, used, research, however, only, can, one, both, each, most, data, when, study, using, such, into, some, number, they, during, where, analysis, there, time, different, high, fish
Estimation models	1124	with, from, as, is, in, of, and, this, for, be, the, to, are, an, at, on, each, not, that, used, or, which, data, was, between, all, also, than, these, more, were, can, two, using, it, number, have, methods, when, but, where, been, fish, both, one, other, however, fisheries, only, if, analysis, their, has, based, because, estimated, such, estimates, different, estimate, use, research, total, some, there, same, size, over, distribution, mean, values, time, then, most, would, into, large, they, new, small, model, could, similar, given, within, study, three, first, those, method
Assessment models	1637	from, an, as, is, in, on, of, and, this, that, for, the, to, be, are, with, not, at, or, have, which, used, it, than, between, also, can, when, these, more, fish, all, where, but, was, however, has, fisheries, other, data, been, two, using, model, research, only, were, such, population, one, each, if, analysis, both, based, values, time, their, some, most, because, different, stock, would, models, there, number, over, management, given, marine, year, size, parameters, into, years, use, methods, first, value, dynamics, mortality, they, assessment, new, biological, then, same, rate, could, estimates, estimated, high, natural, fishery, similar, available, approach, those, should, large, total, its, will, we, species

Appendix 3

Calculating Model Quality

C_V uses four stages to arrive at an overall topic score: (1) segmentation of the topic's top N words into pairs; (2) probability calculations of individual words or pairs of words; (3) calculation of a confirmation measure that captures the agreement of pairs; and finally (4) aggregation of individual confirmation measures into an overall topic coherence score.

(1) The first step is to segment the data into word subsets to calculate the degree of support between two subsets. C_V segments each word in W with every other word in W, where W is the set of a topic's top 15 words. This segmentation creates pairs, S, where the left subset is $W' \in W$ and the right subset is $W^* \in W$. All pairs are formally defined as $S = \{(W', W^*) \mid W' = \{w_i\}; w_i \in W; W^* = W\}$. For example, if $W = \{\text{salmon, catch, tag}\}$, then one pair might be $S_i = (W', W^*)$ as $W' = \{\text{salmon}\}$ and $W^* = \{\text{salmon, catch, tag}\}$.

(2) The probabilities of single words $p(w_i)$ and the joint probability of two words $p(w_i, w_j)$ can be estimated using Boolean document calculation – that is, the number of documents in which w_i or (w_i, w_j) occurs divided by the total number of documents. A Boolean document, however, ignores the frequencies and distances of words. C_v incorporates a Boolean sliding window in which a new virtual document is created for each window of size $s = 110$ (Röder et al., 2015) when sliding over the document, with one word token per step. For example, a document d_1 with w words results in the virtual documents $d_1' = \{w_1, \ldots, w_{110}\}$, $d_2' = \{w_2, \ldots, w_{111}\}$, etc. In contrast to a Boolean document, a Boolean sliding window tries to capture word token proximity to some degree.

(3) For every $S_i = (W', W^*)$ a confirmation measure ϕ is calculated that indicates how strongly W^* supports W' and this confirmation measure is based on the similarity of W' and W^* in relation to all the words in W. To calculate this similarity, W' and W^* are represented as context vectors (Aletras and Stevenson, 2013) as a means to capture the semantic support for all the words in W. These vectors are denoted by $\vec{v}(W')$ and $\vec{v}(W^*)$ and are created by pairing them to all words in W, as exemplified in Equation 3:

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} \text{NPMI}(w_i, w_j)^{\gamma} \right\}_{j=1,\ldots,|W|} \quad (3)$$

Given the running example of $W = \{\text{salmon, catch, tag}\}$, this can be demonstrated with the pair $S_i = (W', W^*)$ as $W' = \{\text{salmon}\}$ and $W^* = \{\text{salmon, catch, tag}\}$. One of these context vectors is $\vec{v}(W') = \vec{v}(salmon)$, now represented as $\vec{v}_{\text{salmon}} = \{\text{NPMI (salmon, salmon)}^{\gamma}, \text{NPMI (salmon, catch)}^{\gamma}, \text{NPMI (salmon, tag)}^{\gamma}\}$.

The coherence between the individual words w_i and w_j is calculated using normalized pointwise mutual information (NPMI), as expressed in Equation 4. In

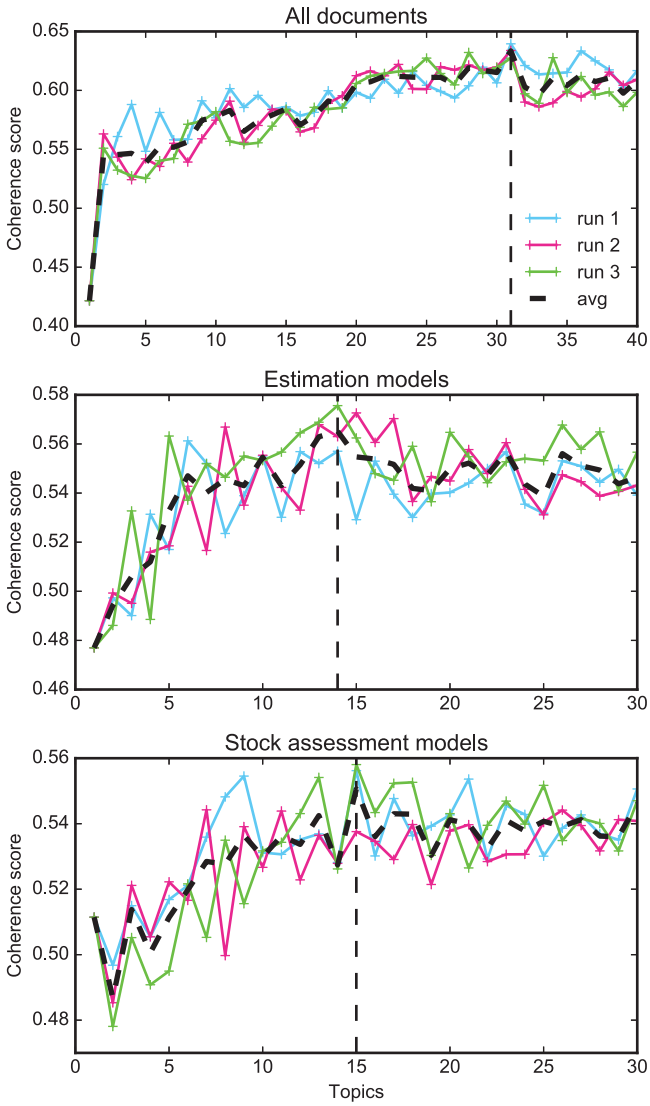


Figure 8. Calculated coherence scores (y-axis) for the number of topics (x-axis) (i.e., *K* parameter) for three different runs. The average coherence score is calculated by averaging the scores over all three runs for the same *K* parameter. The figures represent the following: A: all documents (N = 22,236); B: documents that exhibit the topic estimation models as the dominant topic (N = 1124); C: documents that exhibit the topic stock assessment models as the dominant topic (N = 1637).

contrast to pointwise mutual information (PMI), NPMI shows a higher correlation with human topic ranking data (Bouma, 2009). Additionally, $\varepsilon = 10^{-12}$ (Stevens et al., 2012) is used to account for logarithms of zero, and γ is used to place more weight on higher NPMI values.

$$\mathrm{NPMI}(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i, w_j) + \varepsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j) + \varepsilon)} \right)^{\gamma} \quad (4)$$

Within a pair $S_i = (W', W^*)$, utilizing all context vectors $\vec{v}(W')$, denoted here as $\vec{u}$, and utilizing all context vectors $\vec{v}(W^*)$, denoted here as $\vec{w}$, the cosine vector similarity ϕ_{S_i} is calculated in order to obtain the confirmation measure of the pair $S_i = (W', W^*)$. The cosine vector similarity is expressed in Equation 5.

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \quad (5)$$

(4) Finally, the arithmetic mean of all confirmation measures is taken to obtain the overall coherence score of a topic.

The calculated topic coherence scores can be found in Figure 8.

Appendix 4

See Table 4.

Table 4. The top 15 words (i.e., the words with highest probability) for each of the 31 uncovered general fisheries topics. Topics in bold (i.e., 4 and 9) are the identified modeling topics used in the analysis of this paper, with 4 being the topic estimation models and 9 being the stock assessment models.

Topic	Top 15 words
1	crab, crabs, lobster, eel, eels, size, traps, mm, lobsters, trap, american, anguilla, blue, females, fishery
2	salmon, hatchery, chinook, river, wild, atlantic, survival, coho, sockeye, juvenile, oncorhynchus, fisheries, smolts, pacific, steelhead
3	river, species, sampling, electrofishing, colorado, fishes, chub, capture, population, suckers, sucker, abundance, reach, sites, site
4	**model, estimates, catch, survey, sampling, estimated, models, estimate, distribution, abundance, mean, effort, sample, method, size**
5	genetic, populations, population, river, loci, samples, among, structure, individuals, microsatellite, within, stock, alleles, diversity, sample
6	prey, larvae, growth, larval, food, predation, size, feeding, diet, juvenile, zooplankton, abundance, mm, predator, rates
7	red, reef, gulf, species, snapper, florida, marine, mexico, reefs, fishes, shrimp, coral, habitat, artificial, drum
8	atlantic, bay, striped, tuna, bass, flounder, estuary, north, marine, new, river, carolina, chesapeake, estuaries, estuarine
9	**model, stock, mortality, population, recruitment, models, biomass, year, rate, management, parameters, assessment, fisheries, estimates, fishing**
10	species, variables, environmental, sites, lakes, assemblages, community, water, assemblage, richness, communities, diversity, index, models, spatial
11	cod, sea, atlantic, north, species, herring, size, cm, trawl, length, stock, area, mesh, baltic, fishing
12	fisheries, management, fishing, fishery, catch, economic, marine, effort, fishers, species, recreational, information, anglers, use, new
13	habitat, water, flow, use, depth, river, velocity, substrate, channel, areas, sites, site, area, movement, spawning
14	spawning, females, eggs, egg, males, female, reproductive, male, fecundity, sex, maturity, mature, stage, size, development
15	species, sharks, bycatch, shark, catch, longline, fishery, fisheries, fishing, gear, hooks, caught, hook, cm, atlantic
16	mortality, tag, tagged, tags, tagging, release, survival, released, movement, rates, fisheries, studies, capture, effects, transmitters
17	lake, lakes, perch, michigan, yellow, walleye, great, fisheries, northern, walleyes, mean, ontario, journal, consumption, population
18	growth, length, mm, size, otoliths, body, ages, otolith, cm, mean, years, first, weight, differences, lengths
19	temperature, water, growth, effects, swimming, treatment, energy, levels, temperatures, experiment, activity, body, effect, experiments, experimental
20	sea, squid, mediterranean, distribution, area, anchovy, species, waters, larvae, sardine, marine, spawning, shelf, temperature, mackerel
21	bass, largemouth, reservoir, river, species, lake, catfish, smallmouth, fisheries, shad, water, management, reservoirs, white, black
22	species, fishes, freshwater, carp, new, native, river, water, aquaculture, crayfish, populations, introduced, conservation, tilapia, many
23	species, dna, genetic, gene, mtdna, samples, molecular, mitochondrial, sequence, haplotypes, infection, atlantic, identification, disease, sequences
24	acoustic, depth, vertical, water, bottom, surface, ts, distribution, speed, range, target, density, measurements, night, behaviour
25	otolith, otoliths, sr, marine, river, ratios, samples, water, differences, juvenile, chemistry, isotope, freshwater, values, campana
26	fishing, marine, species, fisheries, areas, sea, area, fishery, catch, australia, effort, total, south, effects, coastal
27	river, sturgeon, dam, chinook, columbia, lower, passage, migration, salmon, downstream, steelhead, upstream, juvenile, spawning, dams
28	trout, brook, rainbow, brown, lake, sea, fry, lamprey, river, stocking, lampreys, salvelinus, arctic, streams, stocked
29	water, concentrations, phytoplankton, production, concentration, samples, nutrient, sediment, carbon, total, food, values, biomass, organic, levels
30	stream, trout, streams, creek, habitat, cutthroat, sites, reaches, river, effects, brook, temperature, watershed, abundance, aquatic
31	sea, pacific, marine, species, climate, ocean, north, alaska, rockfish, change, ecosystem, changes, abundance, temperature, california